# CUNY SPS DATA 621 - CTG5 - HW4

*Betsy Rosalen, Gabrielle Bartomeo, Jeremy O'Brien, Lidiia Tronina, Rose Koh*

*April 24th, 2019*

## Contents

# 1 DATA EXPLORATION

```
## 'data.frame':    8161 obs. of  26 variables:
##  $ INDEX      : int  1 2 4 5 6 7 8 11 12 13 ...
##  $ TARGET_FLAG: int  0 0 0 0 0 1 0 1 1 0 ...
##  $ TARGET_AMT : num  0 0 0 0 0 ...
##  $ KIDSDRIV   : int  0 0 0 0 0 0 0 0 1 0 0 ...
##  $ AGE        : int  60 43 35 51 50 34 54 37 34 50 ...
##  $ HOMEKIDS   : int  0 0 1 0 0 1 0 2 0 0 ...
##  $ YOJ        : int  11 11 10 14 NA 12 NA NA 10 7 ...
##  $ INCOME     : chr  "$67,349" "$91,449" "$16,039" "" ...
##  $ PARENT1    : chr  "No" "No" "No" "No" ...
##  $ HOME_VAL   : chr  "$0" "$257,252" "$124,191" "$306,251" ...
##  $ MSTATUS    : chr  "z_No" "z_No" "Yes" "Yes" ...
##  $ SEX        : chr  "M" "M" "z_F" "M" ...
##  $ EDUCATION  : chr  "PhD" "z_High School" "z_High School" "<High School" ...
##  $ JOB        : chr  "Professional" "z_Blue Collar" "Clerical" "z_Blue Collar" ...
##  $ TRAVTIME   : int  14 22 5 32 36 46 33 44 34 48 ...
##  $ CAR_USE    : chr  "Private" "Commercial" "Private" "Private" ...
##  $ BLUEBOOK   : chr  "$14,230" "$14,940" "$4,010" "$15,440" ...
##  $ TIF        : int  11 1 4 7 1 1 1 1 1 7 ...
##  $ CAR_TYPE   : chr  "Minivan" "Minivan" "z_SUV" "Minivan" ...
##  $ RED_CAR    : chr  "yes" "yes" "no" "yes" ...
##  $ OLDCLAIM   : chr  "$4,461" "$0" "$38,690" "$0" ...
##  $ CLM_FREQ   : int  2 0 2 0 2 0 0 1 0 0 ...
##  $ REVOKED    : chr  "No" "No" "No" "No" ...
##  $ MVR_PTS    : int  3 0 3 0 3 0 0 10 0 1 ...
##  $ CAR_AGE    : int  18 1 10 6 17 7 1 7 1 17 ...
##  $ URBANICITY : chr  "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban/ Urban" "Highly Urban,
```

change BLUEBOOK, HOME_VAL, INCOME $ to numerical value change PARENT1 , Yes-> 2 No ->1 change RED_CAR , yes ->1 no ->0 change SEX into GENDER and if M change into 1 and z_F into 0 split URBANICITY to RURAL and URBAN dummy variables 1,0

Table 1: Data Dictionary

| VARIABLE | DEFINITION | TYPE |
|---|---|---|
| TARGET_FLAG | car crash = 1, no car crash = 0 | response |
| TARGET_AMT | car crash cost = >0, no car crash = 0 | response |
| AGE | driver's age - very young/old tend to be risky | numerical predictor |
| BLUEBOOK | value of vehicle | numerical predictor |
| CAR_AGE | age of vehicle | numerical predictor |
| CAR_TYPE | type of car (6types) | categorical predictor |
| CAR_USE | usage of car (commercial/private) | categorical predictor |
| CLM_FREQ | number of claims past 5 years | numerical predictor |
| EDUCATION | max education level (5types) | categorical predictor |
| HOMEKIDS | number of children at home | numerical predictor |
| HOME_VAL | value of home - home owners tend to drive more responsibly | numerical predictor |
| INCOME | income - rich people tend to get into fewer crashes | numerical predictor |
| JOB | job category (8types, 1missing )- white collar jobs tend to be safer | categorical predictor |
| KIDSDRIV | number of driving children - teenagers likely get into crashes | numerical predictor |
| MSTATUS | maritial status - married people drive more safely | catogerical predictor |
| MVR_PTS | number of traffic tickets | numerical predictor |
| OLDCLAIM | total claims in the past 5 years | numerical predictor |
| PARENT1 | single parent | categorical predictor |
| RED_CAR | a red car | categorical predictor |
| REVOKED | | categorical predictor |
| SEX | gender - woman may have less crashes than man | categorical predictor |
| TIF | time in force - number of years being customer | numerical predictor |
| TRAVTIME | distance to work | numerical predictor |
| URBANCITY | urban/rural | categorical predictor |
| YOJ | years on job - the longer they stay more safe | numerical predictor |

# 2 DATA PREPARATION

# 3 BUILD MODELS

# 4 SELECT MODELS

# 5 Appendix

The appendix is available as script.R file in `project4_insurance` folder.

https://github.com/betsyrosalen/DATA_621_Business_Analyt_and_Data_Mining